PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-331484
(43)Date of publication of application : 30.11.2001

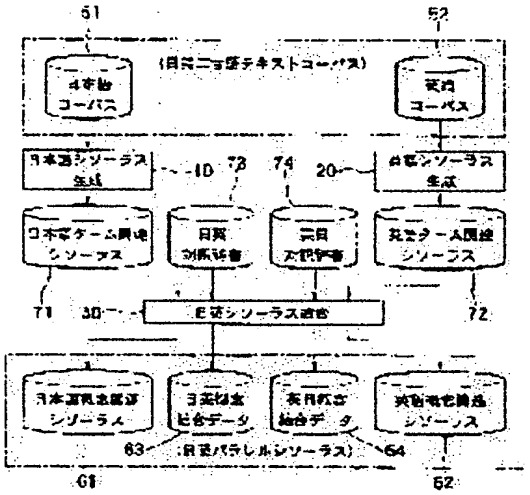
(51)Int.Cl. G06F 17/28
G06F 17/30

(21)Application number : 2000-149413
(22)Date of filing : 22.05.2000

(71)Applicant : HITACHI LTD
(72)Inventor : KAJI HIROYUKI
MORIMOTO YASUTSUGU

(54) RECORDING MEDIUM HAVING PARALLEL THESAURUS GENERATION PROGRAM RECORDED THEREON, RECORDING MEDIUM HAVING PARALLEL THESAURUSES RECORDED THEREON AND RECORDING MEDIUM HAVING PARALLEL THESAURUS NAVIGATION PROGRAM RECORDED THEREON

(57)Abstract:
PROBLEM TO BE SOLVED: To automatically generate parallel thesauruses composed of concept relations within a language and concept connections between the languages and to perform text mining utilizing the generated parallel thesauruses.
SOLUTION: A server computer 1 is provided with Japanese thesaurus generation 10 for generating a Japanese term relation thesaurus 71 by extracting terms from a Japanese corpus 51 and analyzing correlations between the terms, English thesaurus generation 20 for generating English term relation thesaurus 72 by extracting the terms from an English corpus 52 and analyzing the correlations between the terms and Japanese/English thesaurus connection 30 for connecting the Japanese term relation thesaurus 71 and the English term relation thesaurus 72.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－331484
（P2001－331484A）
(43)公開日　平成13年11月30日(2001.11.30)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/28		Ｇ０６Ｆ　17/28	Ｃ	５Ｂ０７５
17/30	１７０	17/30	１７０Ａ	５Ｂ０９１
	３２０		３２０Ｄ	
	３５０		３５０Ｃ	

審査請求　未請求　請求項の数５　ＯＬ　（全 14 頁）

(21)出願番号　特願2000－149413(P2000－149413)

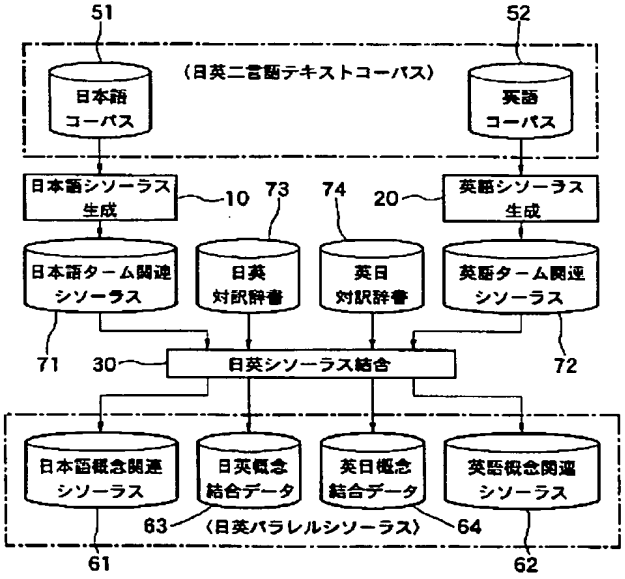
(22)出願日　平成12年５月22日(2000.5.22)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　梶　博行
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　森本　康嗣
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　100091096
弁理士　平木　祐輔

Ｆターム(参考)　5B075 ND03 PQ02 QM07 QP03 UU01
5B091 AA05 AB17 CA12 CB22

(54)【発明の名称】　パラレルシソーラスの生成プログラムを記録した記録媒体、パラレルシソーラスを記録した記録媒体及びパラレルシソーラスナビゲーションプログラムを記録した記録媒体

(57)【要約】
【課題】　言語内の概念関連と言語間の概念結合からなるパラレルシソーラスを自動的に生成する。生成したパラレルシソーラスを利用したテキストマイニングを実現する。
【解決手段】　サーバ計算機１は、日本語コーパス５１からタームを抽出し、ターム間の相関を解析することにより日本語ターム関連シソーラス７１を生成する日本語シソーラス生成１０と、英語コーパス５２からタームを抽出し、ターム間の相関を解析することにより英語ターム関連シソーラス７２を生成する英語シソーラス生成２０と、日本語ターム関連シソーラス７１と英語ターム関連シソーラス７２とを結合する日英シソーラス結合３０とを備える。



【特許請求の範囲】
【請求項１】　コンピュータを、第１言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第１言語のターム関連シソーラスを生成する第１言語シソーラス生成手段と、第２言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第２言語のターム関連シソーラスを生成する第２言語シソーラス生成手段と、第１言語のターム関連シソーラスと第２言語のターム関連シソーラスを結合するシソーラス結合手段と、を備えるパラレルシソーラスの生成装置として機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項２】　前記シーラス結合手段は、対訳辞書を参照して第１言語のターム関連シソーラスと第２言語のターム関連シソーラスとの間で対応するタームを結合するターム結合手段と、結合された第２言語のタームを組み合わせることにより第１言語の各タームから概念ラベルを生成する第１言語概念ラベル生成手段と、結合された第１言語のタームを組み合わせることにより第２言語の各タームから概念ラベルを生成する第２言語概念ラベル生成手段と、言語間のターム結合を言語間の概念結合に変換する概念結合手段と、第１言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第１言語概念関連シソーラス生成手段と、第２言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第２言語概念関連シソーラス生成手段と、同一の第２言語の概念に結合され、関連する第１言語の概念の集合が類似している第１言語の概念をマージして一つの概念にする第１言語概念マージ手段と、同一の第１言語の概念に結合され、関連する第２言語の概念の集合が類似している第２言語の概念をマージして一つの概念にする第２言語概念マージ手段と、を有することを特徴とする請求項１記載のコンピュータ読み取り可能な記録媒体。
【請求項３】　第１言語のタームと第２言語のタームを組み合わせた概念ラベルと概念ラベルが表す概念に基づく概念ラベルの結合とから構成されるシソーラスを記録したことを特徴とするパラレルシソーラスを記録した記録媒体。
【請求項４】　コンピュータを、第１言語の複数のターム又は概念の集合から、第２言語の複数のターム又は概念の集合に遷移する遷移手段を備えるパラレルシソーラスナビゲーションシステムとして機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項５】　前記遷移手段は、第１言語のターム又は概念の集合から、該集合中のターム又は概念に結合された第２言語のターム又は概念に、該第２言語のターム又は概念の関連ターム又は関連概念であって、第１言語のターム又は概念に結合されていないターム又は概念を加えた集合に遷移することを特徴とする請求項４記載のコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、二つの言語のテキストコーパスから二つの言語のシソーラスを結合したパラレルシソーラスを生成する装置及び該パラレルシソーラスを利用したナビゲーションシステムのプログラム、並びにパラレルシソーラスを記録した記録媒体に関する。
【０００２】
【従来の技術】電子化されたテキスト情報の増加と共に、情報アクセス技術の重要性が高まっている。本発明者らは先に特願平１１－２８１０１号として、シソーラスのナビゲーション方法を出願している。この方法は、テキストデータを記憶する文書データベース（以下、テキストコーパスと呼ぶ）から、有益な情報を掘り出す作業（以下、テキストマイニングと呼ぶ）を効率的に行う技術であり、テキストコーパスからターム及びターム間の関連知識を抽出してシソーラスを生成し、該シソーラスの内容をクライアント端末のブラウザに表示することによりナビゲーションする。
【０００３】また、情報処理学会データベースシステム研究会／情報学基礎研究会研究報告ＤＢＳ－１１８－１３／ＦＩ－５４－１３「コーパス対応の関連シソーラスナビゲーション」（１９９９年５月１７日）には、上記ナビゲーション方法に基づくテキストマイニングシステムが報告されている。
【０００４】
【発明が解決しようとする課題】情報検索の分野では、母国語で表現した検索要求を入力して、外国語の文書を検索したいというニーズが高まっている。この情報検索方法は、クロスランゲージ情報検索と呼ばれ、盛んに研究されている。
【０００５】クロスランゲージ情報検索の代表的な手法は、たとえば情報処理学会論文誌４０巻１１号ページ４０７５－４０８６「機械翻訳を用いた英日・日英言語横断検索に関する一考察」（１９９９年１１月）に報告されている。この検索方法は、対訳辞書や機械翻訳システムを利用して、検索要求を文書と同じ言語に翻訳した上で文書検索を実行する。この場合、検索要求は文書に比べて短く文脈情報が少ないため、高精度で翻訳するのが難しく、検索精度が低いという問題がある。
【０００６】この問題に対して、上記特願平１１－２８１０１号で提案しているナビゲーション方法を利用することが考えられる。この場合、単言語から二言語に拡張して情報検索システムのフロントエンドとして使用することにより、クロスランゲージ情報検索における上記問題点を解決することができると考えられるが、二つの言

語のシソーラスを結合することが必要になる。

【０００７】一方、テキストコーパスからのシソーラス生成技術は、上述したようなテキストマイニングへの応用だけでなく、様々な自然言語処理応用システムに有効であるが、次のような技術課題が残されている。

【０００８】従来のシソーラス自動生成の技術は、多義語の取扱いに関して問題がある。従来の技術ではターム間の関連を抽出しているが、ターム間の関連は、本来意味的なものであるので、多義語のタームを語義すなわち概念に分割し、概念間の関連を抽出するのが理想的である。従来の技術によれば、たとえば「bank」の関連タームとして「loan」、「money」、「river」及び「water」等が、「bank」の概念を問わず全て抽出されてしまう。

【０００９】ここで「loan」及び「money」は、お金を預けたり引き出したりする機関としての「bank（銀行）」の関連タームであり、「river」及び「water」は、水辺の場所としての「bank（岸）」の関連タームである。したがって、「bank」をそれが表す概念に分割し、「loan」及び「money」等の関連タームと、「river」及び「water」等の関連タームとが別々に抽出されることが望ましい。

【００１０】また、従来のシソーラス自動生成技術は、同義語の取扱いに関しても問題がある。従来の技術では共起確率、すなわちテキスト中の近傍に揃って出現する確率に基づいて関連タームを抽出している。このため、同義語は関連タームとしてさえ抽出されず、別々のエンティティとして扱われる。同義語は同じ概念を表すのであるから、同義語が一つのエンティティに纏めて取扱われることが望ましい。

【００１１】以上より、本発明の目的は上述した従来の技術における問題点を解決することである。第１の目的は、第１言語の概念及び概念間の関連、第２言語の概念及び概念間の関連、第１言語の概念と第２言語の概念間の結合から構成されるパラレルシソーラスを第１言語及び第２言語のテキストコーパスから自動生成する、パラレルシソーラスの生成プログラムを記録した記録媒体を提供することにある。

【００１２】第２の目的は、クロスランゲージ情報検索のフロントエンドとして、特に、多義語及び同義語の取扱いに注目して、検索要求の高精度な翻訳を可能にする、パラレルシソーラスナビゲーションプログラムを記録した記録媒体を提供することにある。第３の目的は、上記パラレルシソーラスを利用した有効なテキストマイニングを実現するために、パラレルシソーラスを記録した記録媒体を提供することにある。

【００１３】

【課題を解決するための手段】本発明は、コンピュータを、第１言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第１言語のターム関連シソーラスを生成する第１言語シソーラス生成手段と、第２言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第２言語のターム関連シソーラスを生成する第２言語シソーラス生成手段と、第１言語のターム関連シソーラスと第２言語のターム関連シソーラスを結合するシソーラス結合手段と、を備えるパラレルシソーラスの生成装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００１４】また、前記シーラス結合手段は、対訳辞書を参照して第１言語のターム関連シソーラスと第２言語のターム関連シソーラスとの間で対応するタームを結合するターム結合手段と、結合された第２言語のタームを組み合わせることにより第１言語の各タームから概念ラベルを生成する第１言語概念ラベル生成手段と、結合された第１言語のタームを組み合わせることにより第２言語の各タームから概念ラベルを生成する第２言語概念ラベル生成手段と、言語間のターム結合を言語間の概念結合に変換する概念結合手段と、第１言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第１言語概念関連シソーラス生成手段と、第２言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第２言語概念関連シソーラス生成手段と、同一の第２言語の概念に結合され、関連する第１言語の概念の集合が類似している第１言語の概念をマージして一つの概念にする第１言語概念マージ手段と、同一の第１言語の概念に結合され、関連する第２言語の概念の集合が類似している第２言語の概念をマージして一つの概念にする第２言語概念マージ手段と、を有する。これにより、多義語及び同義語が有する概念を考慮したパラレルシソーラスを生成することができる。

【００１５】パラレルシソーラスの生成装置は、以下のように作用する。第１言語シソーラス生成手段は、第１言語のテキストコーパスからターム及び共起するタームの組を抽出し、ターム間の相関を解析することにより、各ターム毎に関連タームの集合を出力する。第１言語が日本語であるとき、たとえば、ターム「銀行」の関連タームの集合として{ローン, 金利, 口座, 利率, 証券, 経済, 金融, 投資}が出力される。

【００１６】第２言語シソーラス生成手段は、第１言語シソーラス生成手段と同様な処理を第２言語のテキストコーパスに対して実行し、各ターム毎に関連タームの集合を出力する。第２言語が英語であるとき、たとえば、ターム「bank」の関連タームの集合として、{account, river, interest, loan, boat, investment, fishing, park, economy, lake}が出力される。

【００１７】シソーラス結合手段において、各手段は以下のように作用する。ターム結合手段は、対訳辞書を参照して、第１言語のターム関連シソーラスと第２言語のターム関連シソーラスとの間で対応するタームを結合す

る。たとえば、「銀行」と「bank」が結合され、「岸」と「bank」が結合される。

【００１８】第１言語概念ラベル生成手段は、結合された第２言語のタームを組み合せることにより、第１言語の各タームから少なくとも１つの概念ラベルを生成する。同様に、第２言語概念ラベル生成手段は、結合された第１言語のタームを組み合せることにより、第２言語の各タームから少なくとも１つの概念ラベルを生成する。たとえば「bank」は「銀行」と「岸」とに結合されている。このとき、「銀行」の関連ターム集合と「岸」の関連ターム集合が似ていなければ、「銀行」と「岸」とが概念的に異なると判断され、「bank」から二つの概念ラベル「bank・銀行」、「bank・岸」が生成される。

【００１９】概念結合手段は、言語間のターム結合を言語間の概念結合に変換する。たとえば、ターム結合「銀行－bank」は概念結合「銀行－bank・銀行」に変換され、ターム結合「岸－bank」は概念結合「岸－bank・岸」に変換される。

【００２０】第１言語概念関連シソーラス生成手段は、第１言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する。同様に、第２言語概念関連シソーラス生成手段は、第２言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する。たとえば、ターム間の関連「bank－interest」は概念間の関連「bank・銀行－interest・金利／利率」に変換され、ターム間の関連「bank－river」は概念間の関連「bank・岸－river」に変換される。

【００２１】第11言語概念マージ手段は、同一の第２言語の概念に結合され、関連する第１言語の概念の集合が類似している第１言語の概念を一つの概念にマージする。同様に、第２言語概念マージ手段は、同一の第１言語の概念に結合され、関連する第２言語の概念の集合が類似している第２言語の概念を一つの概念にマージする。たとえば、日本語の二つの概念「金利」と「利率」が共に英語の概念「interest・金利／利率」に結合されている。このとき、「金利」の関連概念の集合と「利率」の関連概念の集合が類似していれば、「金利」と「利率」とが一つの概念「金利－利率」にマージされる。

【００２２】以上のように各機能が作用することにより、第１言語の概念及び概念間の関連から構成される第１言語の概念関連シソーラスと、第２言語の概念及び概念間の関連から構成される第２言語の概念関連シソーラスとが結合されたパラレルシソーラスが生成される。

【００２３】また、本発明は、第１言語のタームと第２言語のタームを組み合わせた概念ラベルと概念ラベルが表す概念に基づく概念ラベルの結合とから構成されるパラレルシソーラスを記録した記録媒体である。

【００２４】また、本発明は、コンピュータを、第１言語の複数のターム又は概念の集合から、第２言語の複数のターム又は概念の集合に遷移する遷移手段を備えるパラレルシソーラスナビゲーションシステムとして機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００２５】また、前記遷移手段は、第１言語のターム又は概念の集合から、該集合中のターム又は概念に結合された第２言語のターム又は概念に、該第２言語のターム又は概念の関連ターム又は関連概念であって、第１言語のターム又は概念に結合されていないターム又は概念を加えた集合に遷移する。パラレルシソーラスのナビゲーションシステムにおいて、第１言語の概念の集合から第２言語の概念の集合への遷移手段は次のように作用する。

【００２６】遷移前の概念集合中の概念に結合された第２言語の概念と、それらの第２言語の概念と関連が強く、かつ第１言語の概念と結合されていない第２言語の概念を併せて遷移後の概念集合を構成する。これにより、パラレルシソーラスにおいて陽に結合されていない概念を含めて、第１言語の概念集合から関連する第２言語の概念集合へ遷移することができる。

【００２７】

【発明の実施の形態】以下、本発明の実施の形態を添付図面と対応して詳細に説明する。図１は、本発明の実施の形態によるパラレルシソーラスの生成装置と、該装置を含むパラレルシソーラスナビゲーションシステムの構成を説明するブロック図である。なお、本実施の形態では、二つの言語のシソーラスが結合されたパラレルシソーラスとして、日本語と英語とによるパラレルシソーラスの生成について説明する。

【００２８】本実施の形態によるパラレルシソーラスナビゲーションシステム（以下、システムと呼ぶ）は、通信ネットワーク３を介して互いに接続されるサーバ計算機１とクライアント計算機２により構成される。

【００２９】サーバ計算機１は、日本語のシソーラスと英語のシソーラスとを対応付けるパラレルシソーラス生成処理と、本システムの処理のうち、シソーラスの検索処理等を行う。このサーバ計算機１は、処理装置１１、入力装置１２及び記憶装置１３により主に構成される。

【００３０】処理装置１１は、サーバ計算機１の全体の処理を実行する。特に、後述する図２から図４に示すパラレルシソーラスの生成に関わる各データ処理を実行する。入力装置１２は、テキストコーパスの入手媒体に応じてＣＤ－ＲＯＭドライブ、フロッピー（登録商標）ディスクドライブ等であり、テキストコーパスの入力に用いられる。

【００３１】記憶装置１３は、ＲＡＭ、ＲＯＭ、光磁気ディスクライブラリ装置（図示せず）等の記憶手段を総称しており、たとえば、サーバ計算機１の処理プログラム等はＲＯＭに固定的に格納され、サーバ計算機１の処理の過程で作成されたデータ、作業ファイル等はＲＡＭ

に一時的に格納され、さらに各言語のコーパス、シソーラス及び対訳辞書（図２参照）等は光磁気ライブラリ装置等の大容量記憶装置に格納される。

【００３２】クライアント計算機２は、本システムの処理のうち、サーバ計算機１の検索結果として送信されるシソーラスの表示、ユーザとの対話処理等を行う。つぎに、図２から図４を用いてパラレルシソーラスの生成処理の詳細を説明する。

【００３３】図２は、パラレルシソーラス生成装置における入出力データとモジュール構成を機能的に説明する図である。パラレルシソーラス生成装置の入力は、日本語コーパス５１と英語コーパス５２とが対になった日英二言語テキストコーパスである。日本語コーパス５１と英語コーパス５２とには同じ分野のテキストであるという条件が課せられるが、対訳である必要はない。

【００３４】パラレルシソーラス生成装置の出力は、日本語概念関連シソーラス６１、英語概念関連シソーラス６２、日英概念結合データ６３、英日概念結合データ６４からなる日英パラレルシソーラスである。日英概念結合データ６３と英日概念結合データ６４は、情報の内容は同じでレコード形式が違うだけである。冗長ではあるが、日英シソーラス結合処理の効率を考慮して、両方を出力する。

【００３５】パラレルシソーラス生成装置を構成するモジュールは、日本語シソーラス生成１０、英語シソーラス生成２０、及び日英シソーラス結合３０である。日本語シソーラス生成１０は、日本語コーパス５１から日本語ターム関連シソーラス７１を生成する。英語シソーラス生成２０は、英語コーパス５２から英語ターム関連シソーラス７２を生成する。日英シソーラス結合３０は、日英対訳辞書７３と英日対訳辞書７４を参照して、日本語ターム関連シソーラス７１と英語ターム関連シソーラス７２とから日本語概念関連シソーラス６１、英語概念関連シソーラス６２、日英概念結合データ６３、及び英日概念結合データ６４を生成する。日英対訳辞書７３と英日対訳辞書７４は、情報の内容は同じでレコード形式が違うだけである。冗長ではあるが、日英シソーラス結合処理の効率を考慮して、両方を使用する。

【００３６】図３は、日本語シソーラス生成１０の処理の詳細を説明する図である。図３に示すように、日本語シソーラスを生成する処理は、ターム抽出１０１、共起データ抽出１０２、及び相関解析１０３の３つのステップからなる。

【００３７】（１）ターム抽出１０１

日本語コーパス５１からタームを抽出して、出現頻度をカウントする。タームとしては、出現頻度が予め定めた閾値以上の名詞と複合名詞を抽出する。複合名詞は、品詞列パターンを用いたパターンマッチングによって抽出する。高頻度語の中には、特に分野に関係のない一般的な語も多い。それらは、ストップワードリストを用いて取り除く。

【００３８】このストップワードリストに関して、たとえば、“上記”を先頭要素のストップワードとすることにより、“上記システム”というような名詞句を除外できる。同様に、“全体”を末尾要素のストップワードとすることにより、“システム全体”というような名詞句を除外できる。

【００３９】（２）共起データ抽出１０２

共起するタームの対を抽出して、共起頻度をカウントする。共起の定義としてはウィンドウ共起を採用する。すなわち、一定の幅をもったウィンドウをテキストに沿って移動させながら、各位置でのウィンドウに含まれるタームの対を抽出する。ウィンドウの幅は、たとえば機能語を除いて２５タームとする。

（３）相関解析１０３

全てのターム対に対して統計的な相関値を計算し、予め定めた閾値以上の相関値をもつターム対を抽出する。タームの相関値としては相互情報量を用いる。

【００４０】以上述べた（１）～（３）のステップの結果として、日本語ターム関連シソーラス７１が得られる。日本語ターム関連シソーラス７１は、日本語の各タームに対する関連ターム集合を表すレコードの集まりである。すなわち、

$RT_J(x_i)=\{x(i,1), x(i,2), \cdots, x(i,m_i)\}$ （$i=1,2,\cdots,M$）.

ここで、x_iは日本語ターム、$x(i,m_i)$は関連ターム、iは各タームに付される番号、m_iは第i日本語タームの関連ターム数、Mは日本語タームの総数である。レコードの例を以下に示す。

【００４１】RT_J(銀行)＝{ローン, 金利, 口座, 利率, 証券, 経済, 金融, 投資}.

RT_J(金利)＝{ローン, 貸出し, 預貯金, 引き上げ, 銀行}.

以上、日本語シソーラス生成１０の処理を説明したが、英語シソーラス生成２０の処理も、日本語シソーラス生成１０と全く同様である。英語シソーラスの生成処理の結果として、英語ターム関連シソーラス７２が得られる。英語ターム関連シソーラス７２は、英語の各タームに対する関連ターム集合を表すレコードの集まりである。すなわち、

$RT_E(y_i)=\{y(i,1), y(i,2), \cdots, y(i,n_i)\}$ （$i=1,2,\cdots,N$）.

ここで、y_iは英語ターム、$y(i,n_i)$は関連ターム、iは各タームに付される番号、n_iは第i英語タームの関連ターム数、Nは英語タームの総数である。レコードの例を以下に示す。

RT_E(bank)＝{account, river, interest, loan, boat, investment, fishing, park, economy, lake}.

RT_E(interest)＝{loan, deposit, bank, sciene, economy, exchange, politics}.

【００４２】図４は、日英シソーラス結合モジュール３０の処理の詳細を説明する図である。図４に示すように、日英シソーラス結合処理は、日英ターム結合３０１、日本語概念ラベル生成３０２、英語概念ラベル生成３０３、日英概念結合３０４、日本語概念関連シソーラス生成３０５、英語概念関連シソーラス生成３０６、日本語概念マージ３０７、及び英語概念マージ３０８の８つのステップからなる。以下、これらの処理の詳細を説明する。

【００４３】（１）日英ターム結合３０１

日英対訳辞書７３と英日対訳辞書７４を参照して、日本語ターム関連シソーラス７１と英語ターム関連シソーラス７２の間で対応するタームを結合し、日英ターム結合データ９１と英日ターム結合データ９２とを出力する。日英ターム結合の入力のうち、日本語ターム関連シソーラス７１と英語ターム関連シソーラス７２とは既に説明したので、日英対訳辞書７３と英日対訳辞書７４とについて説明する。

【００４４】日英対訳辞書７３は、日本語の各タームに対する対訳英語ターム集合を表すレコードの集まりである。すなわち、

$D_{JE}(a_i)=\{b(i,1), b(i,2), \cdots, b(i,l_i)\}$　（$i=1, 2, \cdots, K$）.

ここで、ａは日本語ターム、ｂは英語タームである。レコードの例を以下に示す。

D_{JE}(銀行)＝{bank}.

D_{JE}(岸)＝{bank}.

【００４５】英日対訳辞書７４は、英語の各タームに対する対訳日本語ターム集合を表すレコードの集まりである。すなわち、

$D_{EJ}(b_i)=\{a(i,1), a(i,2), \cdots, a(i,k_i)\}$　（$i=1, 2, \cdots, L$）.

ここで、ｂは英語ターム、ａは日本語タームである。レコードの例を以下に示す。

D_{EJ}(bank)＝{銀行, バンク, 岸}.

D_{EJ}(interest)＝{興味, 金利, 利率}.

【００４６】次に、日英ターム結合の出力について説明する。日英ターム結合データ９１と英日ターム結合データ９２とは、同じ情報を異なる形式で表現したものである。冗長ではあるが、後続の処理の効率を考慮して両方を出力する。日英ターム結合データ９１は、日本語タームの各々について、それに結合された英語タームの集合を表すレコードの集まりである。すなわち、

$TL_{JE}(x_i)=\{y'(i,1), y'(i,2), \cdots, y'(i,n'_i)\}$

（$i=1, 2, \cdots, M$）.

ここで、ｘは日本語ターム，ｙ'は英語タームである。

【００４７】英日ターム結合データ９２は、英語タームの各々について、それに結合された日本語タームの集合を表すレコードの集まりである。すなわち、

$TL_{EJ}(y_i)=\{x'(i,1), x'(i,2), \cdots, x'(i,m'_i)\}$

（$i=1, 2, \cdots, N$）.

ここで、ｙは英語ターム、ｘ'は日本語タームである。

【００４８】日英ターム結合３０１のアルゴリズムは次のとおりである。

１）日英ターム結合データ９１を初期化する。すなわち、

$TL_{JE}(x_i) \leftarrow \phi$　（$i=1, 2, \cdots, M$）.

２）英日ターム結合データ９２を初期化する。すなわち、

$TL_{EJ}(y_i) \leftarrow \phi$　（$i=1, 2, \cdots, N$）.

【００４９】３）次の２つの条件を満足する、日本語タームｘと英語タームｙとを結合する。

（ａ）対訳関係＜ｘ，ｙ＞が対訳辞書によってサポートされている。

（ｂ）対訳関係＜ｘ，ｙ＞のドメイン関連度ＤＲ（ｘ，ｙ）が予め定めた閾値以上である。

すなわち、（ａ）及び（ｂ）を満足する全ての日本語タームｘと英語タームｙの対に関して、

$TL_{JE}(x) \leftarrow TL_{JE}(x) \cup \{y\}$及び$TL_{EJ}(y) \leftarrow TL_{EJ}(y) \cup \{x\}$.

を実行する。

（ａ）ｘがｋ個のターム$x_1, x_2, \cdots, x_k$の並び，ｙがｋ個のターム$y_1, y_2, \cdots, y_k$の並びであって、$\{y'_1, y'_2, \cdots, y'_k\}=\{y_1, y_2, \cdots, y_k\}$であるような$y'_1(\in D_{JE}(x_1)), y'_2(\in D_{JE}(x_2)), \cdots, y'_k(\in D_{JE}(x_k))$が存在する。

（ｂ）　$DR(x, y) \geqq \theta$.

【００５０】条件（ａ）により、日本語タームｘと英語タームｙとが対訳関係にあるかを知るために、構成要素の間での対訳関係が成立しているかが、集合におけるタームの順番を問わずにチェックされる。特に、ｋ＝１のときｙ∈D_{JE}(ｘ)となる。すなわち、対訳辞書に登録されているターム対であることを意味する。ｋ≧２のとき、構成要素間の対訳関係が、対訳辞書に登録されているような複合語タームの対であることを意味する。条件（ｂ）により、対訳辞書が示唆する対訳関係がドメインで成立する関係であるかがチェックされる。対訳関係＜ｘ，ｙ＞のドメイン関連度ＤＲ（ｘ，ｙ）は次式で定義される。

【００５１】

【数１】

$$DR(x,y)=max\{DR_{JE}(x,y),DR_{EJ}(x,y)\}.$$

$$DR_{JE}(x,y)=\left|\bigcup_{x'\in RT_J(x)}D_{JE}(x')\cap RT_E(y)\right|/\left|RT_J(x)\right|.$$

$$DR_{EJ}(x,y)=\left|\bigcup_{y'\in RT_E(y)}D_{EJ}(y')\cap RT_J(x)\right|/\left|RT_E(y)\right|.$$

【００５２】$DR_{JE}(x,y)$は、日本語タームの関連タームのうち、英語訳が英語タームの関連タームであるものの比率である。すなわち、ある日本語タームがどのような文脈で出現するかを示す出現文脈が、英語タームの出現文脈と重なる度合である。また、$DR_{EJ}(x,y)$は、英語タームの出現文脈が日本語タームの出現文脈と重なる度合である。多少なりとも出現文脈に共通性があれば、対訳関係がドメインで成立すると考えてよいので、ドメイン関連度の閾値θは小さめに設定するのがよい。以上のアルゴリズムにより、日英ターム結合３０１が実行される。

【００５３】（２）日本語概念ラベル生成３０２
（３）英語概念ラベル生成３０３
日本語概念ラベル生成３０２と英語概念ラベル生成３０３の処理は、日本語と英語の役割が反転する以外は全く同様である。したがって、ここでは英語概念ラベル生成３０３について説明する。

【００５４】英語概念ラベル生成ステップ３０３は、英日ターム結合データ９２と日本語ターム関連シソーラス７１とに基づいて、英語タームの各々から一つ以上の英語概念ラベルを生成する。さらに、生成した英語概念ラベルの各々に対して関連タームの集合を生成する。このために、英語ターム関連シソーラス７２、日英ターム結合データ９１、日本語ターム関連シソーラス７１、英日ターム結合データ９２を参照する。

【００５５】英語概念ラベル生成３０３の入力データは既に説明済みであるので、出力データを説明する。英語概念ラベルは、タームを組み合せたものであり、以下のように定義される。
＜英語概念ラベル＞：＝＜英語ターム＞｜＜英語ターム＞・＜日本語修飾子＞｜＜英語概念ラベル＞+＜英語概念ラベル＞.
＜日本語修飾子＞：＝＜日本語ターム＞|＜日本語修飾子＞/＜日本語ターム＞.

【００５６】＜英語ターム＞・＜日本語ターム＞/.../＜日本語ターム＞は、英語タームが表す概念のうち、日本語タームが表す概念と共通の概念を指示する。たとえば、「bank･銀行」は、お金に関わる業務を行う組織としてのbankを指示し、「bank･岸」は、川や湖に沿った場所としてのbankを指示する。＜英語概念ラベル＞+＜英語概念ラベル＞は、二つの概念ラベルが指示する概念の共通部分を核とし、それぞれの概念ラベルが指示する概念を合わせた範囲の概念を指示する。「duty･税+tax」、「plane･飛行機+airplane」、及び「reasoning+inference」などが例である。

【００５７】英語概念ラベルデータ９４は、各英語タームｙに対応する英語概念ラベル集合を表すレコード$C_E(y)$の集まりである。英語概念ラベル集合の例を示す。
C_E(interest)＝{interest･興味, interest･金利/利率}.
このレコードC_E(interest)は、英日ターム結合データ９２中に
TL_{EJ}(interest)＝{興味, 金利, 利率}.
なるレコードが含まれるとき、それに対応して生成される。

【００５８】英語概念の関連タームデータ９６は、各英語概念ラベルに対する関連ターム集合を表すレコードの集まりで、以下のように記す。
$RT_E(Y_i)=\{y(N+i,1),y(N+i,2),\cdots,y(N+i,n_{N+i})\}$
$(i=1,2,\cdots,Q)$.
ここで、Ｙは英語概念ラベル、ｙは英語タームである。

【００５９】英語概念の関連タームデータ９６は、最終目的である英語概念関連シソーラス６２を生成するための中間データである。英語概念関連シソーラス６２として必要なのは、関連ターム集合ではなく、関連概念集合である。しかし、全てのタームに対する概念ラベル集合を生成してからでないと、関連概念集合を作成することはできない。そこで、暫定的に関連ターム集合を作成しておき、後続の英語概念関連シソーラス生成３０６において関連概念集合に変換する。

【００６０】英語タームｙに対する英語概念ラベル集合$C_E(y)$と、$C_E(y)$中の概念ラベルに対する関連ターム集合を生成するアルゴリズムは次のとおりである。
１）英語タームｙが少なくとも一つの日本語タームと結合されているとき
ｉ）英語概念ラベル集合の初期データを作成する。英語タームｙに結合された日本語タームの各々を日本語修飾子とする英語概念ラベルを生成し、その要素とする。すなわち、
$C_E(y)\leftarrow\{y\cdot x\mid x\in TL_{EJ}(y)\}$.
ii）二つの英語概念の類似度が予め定めた閾値α以上であるなら、それらを一つの英語概念に統合する処理を可能な限り繰り返す。すなわち、

While $\exists\, y\cdot x_1/x_2/\cdots/x_k(\in C_E(y))$
and $y\cdot x'_1/x'_2/\cdots/x'_{k'}(\in C_E(y))$

$$\text{s.t. } S(y\cdot x_1/x_2/\cdots/x_k,\ y\cdot x'_1/x'_2/\cdots/x'_{k'})\geqq\alpha,$$
$$C_E(y)\leftarrow C_E(y)-\{y\cdot x_1/x_2/\cdots/x_k,\ y\cdot x'_1/x'_2/\cdots/x'_{k'}\}$$
$$+\{y\cdot x_1/x_2/\cdots/x_k/x'_1/x'_2/\cdots/x'_{k'}\}.$$

ここで、共通の英語タームｙに関わる二つの英語概念Ｙ　　【００６１】

$_1=y\cdot x_1/x_2/\cdots/x_k$とＹ$_2=y\cdot x'_1/x'_2/\cdots/x'_{k'}$　　【数２】

の類似度Ｓ（Ｙ$_1$，Ｙ$_2$）は次式で定義される。

$$S(Y_1,Y_2)=|(\bigcup_{1\leq i\leq k}RT_J(x_i))\cap(\bigcup_{1\leq i\leq k'}RT_J(x'_i))|$$
$$/|(\bigcup_{1\leq i\leq k}RT_J(x_i))\cup(\bigcup_{1\leq i\leq k'}RT_J(x'_i))|.$$

【００６２】すなわち、概念ラベルを構成する日本語修飾子の関連ターム集合間の重なり度で定義される。

iii）処理ii）の結果として得られた英語概念の各々に対して関連ターム集合データを作成する。英語概念ｙ・$x_1/x_2/\cdots/x_k$の関連ターム集合$RT_E(y\cdot x_1/x_2/\cdots/x_k)$は次式のとおりである。　　【００６３】　　【数３】

$$RT_E(y\cdot x_1/x_2/\cdots/x_k)=RT_E(y)$$
$$-\{\bigcup_{x''\in\bigcup_{i\in JM2}RT_J(x_i)}TL_{JE}(x'')-\bigcup_{x''\in\bigcup_{i\in JM1}RT_J(x_i)}TL_{JE}(x'')\}.$$

【００６４】ここで、ＪＭ１は日本語修飾子の要素の集合である。すなわち、ＪＭ１＝｛x_1，x_2，…，x_k｝．ＪＭ２は英語タームｙに結合された日本語タームの集合から日本語修飾子の要素を除いたものである。すなわち、ＪＭ２＝$TL_{EJ}(y)-JM1$．

２）英語タームｙが日本語タームｘと結合されていないとき

ｉ）英語タームｙそのものを概念ラベルとする。英語概念ラベル集合はこれを唯一の要素とする。すなわち、$C_E(y)\leftarrow\{y\}$．

ii）英語タームｙの関連ターム集合$RT_E(y)$をそのまま英語概念ラベルｙの関連ターム集合とする。

上記アルゴリズムの１）のii）における閾値αの設定について補足しておく。ここでの目的は、一つの英語タームが表す複数の概念を区別するための日本語修飾子を得ることである。したがって、閾値αは小さめに設定し、類義の日本語訳語を一つの日本語修飾子に統合するのがよい。

【００６５】日本語概念ラベル生成３０２は、英語概念ラベル生成３０３と同様である。その出力である日本語概念ラベルデータ９３は、英語概念ラベルデータ９４と同様で、日本語の各タームｘに対応する日本語概念ラベル集合を表すレコード$C_J(x)$の集まりである。日本語概念ラベルは英語概念ラベルと同様で、以下のようにタームを組み合わせたものである。

【００６６】＜日本語概念ラベル＞：＝＜日本語ターム＞|＜日本語ターム＞・＜英語修飾子＞|＜日本語概念ラベル＞+＜日本語概念ラベル＞．

＜英語修飾子＞：＝＜英語ターム＞|＜英語修飾子＞/＜英語ターム＞．

日本語概念ラベル生成３０２のもう一つの出力である日本語概念の関連タームデータ９５は、英語概念の関連タームデータ９６と同様で、日本語の各概念ラベルに対する関連ターム集合を表すレコードの集まりである。すなわち、

$RT_J(X_i)=\{x(M+i,1),x(M+i,2),\cdots,x(M+i,n_{M+i})\}$
（ｉ＝１，２，…，Ｐ）．

ここで、Ｘは日本語概念ラベル、ｘは日本語タームである。

【００６７】（４）日英概念結合３０４

日英概念結合３０４は、日本語概念ラベルデータ９３と英語概念ラベルデータ９４とを入力として、日英概念結合データ６３と英日概念結合データ６４とを生成する。日本語概念ラベルデータ９３と英語概念ラベルデータ９４とは説明済みであるので、まず日英概念結合データ６３と英日概念結合データ６４とについて説明する。

【００６８】日英概念結合データ６３は、日本語の各概念について、それに結合された英語の概念の集合を表すレコードの集まりである。すなわち、

$CL_{JE}(X_i)=\{Y'(i,1),Y'(i,2),\cdots,Y'(i,q'_i)\}$
（ｉ＝１，２，…，Ｐ）．

ここで、Ｘは日本語概念ラベル、Ｙ'は英語概念ラベルである。

【００６９】同様に、英日概念結合データ６４は、英語の各概念について、それに結合された日本語の概念の集合を表すレコードの集まりである。すなわち、

$CL_{EJ}(Y_i)=\{X'(i,1),X'(i,2),\cdots,X'(i,p'_i)\}$
（ｉ＝１，２，…，Ｑ）．

ここで、Ｙは英語概念ラベル、Ｘ'は日本語概念ラベルである。

【００７０】日英概念結合データ６３と英日概念結合データ６４とを生成するアルゴリズムは次のとおりであ

る。

1）全ての日本語概念ラベル$X=x\cdot y_1/y_2/\cdots/y_{k'}$に対して、

$$CL_{JE}(X)=\{Y\mid Y=y\cdot x_1/x_2/\cdots/x_k(\in C_E(y)),\ y\in\{y_1,y_2,\cdots,y_{k'}\},\ \{x_1,x_2,\cdots,x_k\}\ni x\}.$$

すなわち、Xの英語修飾子に含まれる英語タームyの英語概念集合$C_E(y)$の要素Yであって、日本語修飾子にxを含むものの集合を生成する。

$$CL_{EJ}(Y)=\{X\mid X=x\cdot y_1/y_2/\cdots/y_k(\in C_J(x)),\ x\in\{x_1,x_2,\cdots,x_{k'}\},\ \{y_1,y_2,\cdots,y_k\}\ni y\}.$$

すなわち、Yの日本語修飾子に含まれる日本語タームxの日本語概念集合$C_J(x)$の要素Xであって、英語修飾子にyを含むものの集合を生成する。

【００７２】日英概念結合３０４の出力例を示す。日本語概念ラベルデータ９３が

C_J(興味)＝{興味･interest}、

C_J(金利)＝{金利･interest}、

C_J(利率)＝{利率･interest}

であり、英語概念ラベルデータ９４が

C_E(interest)＝{interest･興味, interest･金利/利率}

であるとする。このとき、日英概念結合データ６３として

CL_{JE}(興味･interest)＝{interest･興味}、

CL_{JE}(金利･interest)＝{interest･金利/利率}、

CL_{JE}(利率･interest)＝{interest･金利/利率}

が生成され、英日概念結合データ６４として

CL_{EJ}(interest･興味)＝{興味･interest}、

CL_{EJ}(interest･金利/利率)＝{金利･interest, 利率･interest}

が生成される。

【００７３】ここで、概念ラベルの表記に関する一つの規則を定める。ある英語タームに対応する英語概念がただ一つであるならば、日本語修飾子をつけることは無意味であり、タームそのものを概念ラベルとして差し支えない。日本語タームに関しても同様である。上に述べた日英概念結合３０４のアルゴリズムと違って、これ以降の処理では、日本語修飾子や英語修飾子に基づく判断を含まない。したがって、タームの唯一の概念である概念については、概念結合データ６３，６４の出力時に、概念ラベルをタームそのものに変更することにする。この規則に従えば、上の例における日英概念結合データ６３は

CL_{JE}(興味)＝{interest･興味}、

CL_{JE}(金利)＝{interest･金利/利率}、

CL_{JE}(利率)＝{interest･金利/利率}

となり、英日概念結合データ６４は

CL_{EJ}(interest･興味)＝{興味}、

CL_{EJ}(interest･金利/利率)＝{金利, 利率}

となる。ここでは、「興味」が単一の概念を表す語であるので、概念ラベル「興味･interest」が「興味」に略記されている．「金利･interest」「利率･interest」も同様で、それぞれ「金利」「利率」に略記されている．一方、「interest」は複数の概念を表す語であるので、概念ラベル「interest･興味」「interest･金利/利率」を略記することはできない。

【００７１】2）全ての英語概念ラベル$Y=y\cdot x_1/x_2/\cdots/x_{k'}$に対して、

【００７４】（５）日本語概念関連シソーラス生成３０５

（６）英語概念関連シソーラス生成３０６

日本語概念関連シソーラス生成３０５は、日本語概念ラベルデータ９３と日本語概念の関連タームデータ９５とを入力して、日本語概念関連シソーラス６１を出力する。英語概念関連シソーラス生成３０６は、英語概念ラベルデータ９４と英語概念の関連タームデータ９６とを入力として、英語概念関連シソーラス６２を出力する。これらの入力については説明済みである。

【００７５】出力である日本語概念関連シソーラス６１と英語概念関連シソーラス６２は次のとおりである。日本語概念関連シソーラス６１は、日本語の各概念の関連概念集合を表すレコードの集まりである。すなわち、

$RC_J(X_i)=\{X(i,1),X(i,2),\cdots,X(i,p_i)\}$　（i＝1，2，…，P）．

ここで、Xは日本語概念ラベルである。関連概念集合の例を以下に示す。

【００７６】RC_J(銀行)＝{ローン, 金利, 口座, 利率, 証券, 経済, 金融, 投資}．

RC_J(岸)＝{川, 水, ボート, 湖, 釣り}．

英語概念関連シソーラス６２は、英語の各概念の関連概念集合を表すレコードの集まりである。すなわち、

$RC_E(Y_i)=\{Y(i,1),Y(i,2),\cdots,Y(i,q_i)\}$　（i＝1，2，…，Q）．

ここで、Yは英語概念ラベルである。関連概念集合の例を以下に示す。

RC_E(bank･銀行)＝{account･口座, interest･金利/利率, loan, investment, economy}．

RC_E(bank･岸)＝{river, boat, water, fishing, park･公園, lake}．

【００７７】英語概念関連シソーラス生成３０６のアルゴリズムは以下のとおりである。なお、日本語概念関連

シソーラス生成３０５のアルゴリズムも全く同様である。英語概念Ｙの関連ターム集合$RT_E(Y)$の各要素ｙに対応して、ｙの概念ラベル集合$C_E(y)$の要素のうちＹとの相関度が最大のものを関連概念集合$RC_E(Y)$の要素として選択する。すなわち、

【００７８】

【数４】

$$RC_E(Y)=\{Y' \mid y\in RT_E(Y), Y'\in C_E(y), S_2(Y', Y) = \max_{Y''\in C_E(y)} S_2(Y'', Y)\}.$$

【００７９】ここで、S_2は関連ターム集合に基づく英語概念の相関度で、次式で定義される。

$S_2(Y_1, Y_2)=|RT_E(Y_1)\cap RT_E(Y_2)|/|RT_E(Y_1)\cup RT_E(Y_2)|$.

たとえば、英語概念「bank･銀行」の関連ターム集合が「interest」を含み、英語ターム「interest」の概念ラベル集合が{interest･興味, interest･金利/利率}であるとする。このとき、「bank･銀行」と「interest･興味」との相関度、「bank･銀行」と「interest･金利/利率」との相関度が計算される。後者の相関度が大きければ、「bank･銀行」の関連概念集合の要素として「interest･金利/利率」が選択される。

【００８０】（７）日本語概念マージ３０７

（８）英語概念マージ３０８

日本語概念マージ３０７と英語概念マージ３０８の処理は、日本語と英語の役割が反転する以外、全く同様である。したがって、ここでは英語概念マージ３０８について説明する。

【００８１】英語概念マージ３０８は、日本語の同一概念に結合された英語概念で類似度の高いものをマージして一つの概念にする。入力は、英語概念関連シソーラス６２、日英概念結合データ６３、英日概念結合データ６４であり、出力はそれらの更新データである。

【００８２】英語概念マージ３０８のアルゴリズムは以下のとおりである。全ての日本語概念Ｘに関して、以下の処理を可能な限り繰り返す。$Y_1, Y_2\in CL_{JE}(X)$で$S_3(Y_1, Y_2)\geqq\beta$なる英語概念の組Y_1, Y_2が存在するならば、ａ）からｃ）を実行する。ここで、$S_3(Y_1, Y_2)$は英語概念Y_1とY_2の類似度で、次式で定義される。

$S_3(Y_1, Y_2)=|RC_E(Y_1)\cap RC_E(Y_2)|/|RC_E(Y_1)\cup RC_E(Y_2)|$.

【００８３】ａ）英語概念関連シソーラス６２の更新

全ての$Y\in RC_E(Y_1)$に関して、$RC_E(Y)\leftarrow RC_E(Y)-\{Y_1\}+\{Y_1+Y_2\}$.

全ての$Y\in RC_E(Y_2)$に関して、$RC_E(Y)\leftarrow RC_E(Y)-\{Y_2\}+\{Y_1+Y_2\}$.

$RC_E(Y_1+Y_2)\leftarrow RC_E(Y_1)\cup RC_E(Y_2)$.

$RC_E(Y_1)$と$RC_E(y_2)$を消去する。

【００８４】ｂ）日英概念結合データ６３の更新

全ての$x\in CL_{EJ}(Y_1)$に関して、$CL_{JE}(X)\leftarrow CL_{JE}(X)-\{Y_1\}+\{Y_1+Y_2\}$.

全ての$X\in CL_{EJ}(Y_2)$に関して、$CL_{JE}(X)\leftarrow CL_{JE}(X)-\{Y_2\}+\{Y_1+Y_2\}$.

ｃ）英日概念結合データ６４の更新

$CL_{EJ}(Y_1+Y_2)\leftarrow CL_{EJ}(Y_1)\cup CL_{EJ}(Y_2)$.

$CL_{EJ}(Y_1)$と$CL_{EJ}(Y_2)$を消去する。

【００８５】上記アルゴリズム中の閾値βは大きめに設定し、類似度が非常に高い概念のみをマージするのがよい。概念の範囲やニュアンスが異なるタームを別々のエンティティとするほうが、利用価値の高いシソーラスになるからである。この点は、相手言語のタームが表す複数の概念を区別することが目的の場合（英語概念ラベル生成３０３のアルゴリズムにおける閾値α）と異なっている。

【００８６】以上説明した処理によって、日本語コーパス５１と英語コーパス５２が対になった日英二言語テキストコーパスから、日本語概念関連シソーラス６１、英語概念関連シソーラス６２、日英概念結合データ６３、英日概念結合データ６４からなる日英パラレルシソーラスを生成することができる。このように生成された日英パラレルシソーラスは、図１に示す通信ネットワーク３を介して、クライアント計算機２によるシソーラスナビゲーションに利用される。つぎに、図５～図７を用いて本システムの処理を説明する。

【００８７】図５は、本システムにおいて、クライアント計算機２の表示画面の内容を説明する図である。図５に示す表示画面は、概念集合エリア１０１０、ズームインエリア１０２０及び機能選択ボタンから構成される。機能選択ボタンには、ズームインボタン１０３０、翻訳ボタン１０４０、クリアボタン１０５０、終了ボタン１０６０がある。

【００８８】ズームインエリア１０２０には、一つ以上の概念クラスタ１０２１がそれに対応付けられた選択ボタン１０２２とともに表示される。この概念クラスタ１０２１は、関連性の高い概念の集合である。たとえば、“地球環境問題”に該当する概念クラスタであれば、「地球温暖化」、「オゾン層」、「温室効果」、「フロン」及び「大気」等の概念が表示される。クライアント計算機２のユーザは、これら概念クラスタ１０２１を複数指定することができる。

【００８９】図６は、本実施の形態によるパラレルシソーラスナビゲーションシステムの処理を説明するフローチャートである。以下、図５に示した表示内容と対応して本システムの処理を説明する。最初に初期画面を表示

する（ステップ４１０）。初期画面では、概念集合エリア１０１０とズームインエリア１０２０は空白である。本システムには、「日本語」／「英語」を切り替える言語インジケータが内部に設けられており、初期画面を表示したときには、言語インジケータを「日本語」にする。

【００９０】初期画面表示（ステップ４１０）の後、入力待ちの状態になる（ステップ４２０）。この状態では、概念集合エリア１０１０は書き込み可能であり、通常、ユーザが一つ以上の日本語タームあるいは日本語概念ラベルを書き込む。入力待ちの状態で押されたボタンにより、以下のように分岐する。

【００９１】（１）ズームインボタン１０３０が押されたとき

概念集合エリア１０１０に表示されている概念集合を読み込む（ステップ４３０）。ユーザが概念集合エリア１０１０に書き込むのは、通常、概念ラベルでなくタームである。タームが書き込まれている場合には、該タームから生成された全ての概念ラベルが書き込まれているとみなして処理する。この処理は、言語インジケータが「日本語」のときには日本語概念関連シソーラス６１を参照し、言語インジケータが「英語」のときには英語概念関連シソーラス６２を参照することにより行われる。

【００９２】つぎに、概念集合に含まれる概念と関連の強い概念を加えて概念集合を拡大し、拡大された概念集合をクラスタリングする（ステップ４４０）。この処理は、言語インジケータが「日本語」のときには日本語概念関連シソーラス６１を参照し、言語インジケータが「英語」のときには英語概念関連シソーラス６２を参照することにより行われる。最後に、得られた概念クラスタを選択ボタン１０２２とともにズームインエリア１０２０に表示し（ステップ４５０）、入力待ちの状態に戻る。

【００９３】（２）概念クラスタの選択ボタン１０２２が押されたとき

選択された概念クラスタ１０２１を概念集合エリア１０１０にコピー（上書き）し（ステップ４６０）、入力待ちの状態に戻る。入力待ちの状態では、概念集合エリア１０１０は書き込み可能であり、ユーザがタームあるいは概念ラベルを追加したり、削除したりすることが可能である。

【００９４】（３）翻訳ボタン１０４０が押されたとき

概念集合エリア１０１０に表示されている概念集合を読み込む（ステップ４７０）。この処理はステップ４３０と全く同じである。つぎに、概念集合を翻訳する（ステップ４８０）。言語インジケータが「日本語」のときには日本語概念集合から英語概念集合への翻訳が実行され、言語インジケータが「英語」のときには英語概念集合から日本語概念集合への翻訳が実行される。

【００９５】最後に、翻訳結果を概念集合エリア１０１０に表示（上書き）し、言語インジケータを反転させ（ステップ４９０）、入力待ちの状態に戻る。入力待ちの状態では、概念集合エリア１０１０は書き込み可能であり、ユーザがタームあるいは概念ラベルを追加したり、削除したりすることが可能である。

【００９６】（４）クリアボタン１０５０が押されたとき

初期画面表示状態（ステップ４１０）に戻る。

（５）終了ボタン１０６０が押されたとき

処理を終了する。

以上述べた処理により、言語間の遷移を含むパラレルシソーラスのナビゲーションが可能になる。

【００９７】図７は、本システムを特徴付ける概念集合翻訳（ステップ４８０）の処理を詳細に説明する図である。図７は、日本語概念集合を英語概念集合に翻訳する処理を示したものであるが、英語概念集合を日本語概念集合に翻訳する処理も全く同様である。

【００９８】入力日本語概念集合が与えられると、日英概念結合データ６３を参照して、日本語概念集合中の日本語概念に結合されている英語概念を集めて、核となる英語概念集合を生成する（ステップ４８１）。

【００９９】つぎに、英語概念関連シソーラス６２を参照して、この英語概念集合に含まれる英語概念と関連の強い英語概念を集め、さらに、英日概念結合データ６４を参照して、関連英語概念のうちで日本語概念と結合されていないものを選択する。核となる英語概念集合に選択した英語概念を追加して翻訳結果とする（ステップ４８２）。

【０１００】日本語概念集合から英語疑念集合への翻訳（遷移）の例を示す。入力の日本語概念集合は｛地球温暖化，オゾン層，温室効果，フロン，大気，二酸化炭素，環境｝であるとする。日英概念結合データ６３は以下のレコードを含むとする。

【０１０１】CL_{JE}(地球温暖化)＝φ.

CL_{JE}(オゾン層)＝{ozone layer}.

CL_{JE}(温室効果)＝φ.

CL_{JE}(フロン)＝φ.

CL_{JE}(大気)＝{atmosphere･大気}.

CL_{JE}(二酸化炭素)＝{carbon dioxide}.

CL_{JE}(環境)＝{environment}.

さらに、英語概念関連シソーラス６２が以下のレコードを含むとする。

【０１０２】RC_J(ozone layer)＝{chrolofluorocarbon, depletion, atmosphere･大気, warming}.

RC_J(atmosphere･大気)＝{pollution, environment, gas･気体/ガス, carbon dioxide}.

RC_J(carbon dioxide)＝{atmosphere･大気, energy, warming, environment, regulation}.

RC_J(environment)＝{protection, carbon dioxide, energy, atmosphere･大気, pollution}.

また、英日概念結合データ６４は以下のレコードを含むとする。

【０１０３】CL_{EJ}(ozone layer)＝{オゾン層}.
CL_{EJ}(chrolofluorocarbon)＝φ.
CL_{EJ}(depletion)＝{破壊}.
CL_{EJ}(atmosphere･大気)＝{大気}.
CL_{EJ}(warming)＝φ.
CL_{EJ}(pollution)＝{汚染}.
CL_{EJ}(environment)＝{環境}.
CL_{EJ}(gas･気体/ガス)＝{気体, ガス}.
CL_{EJ}(carbon dioxide)＝{二酸化炭素}.
CL_{EJ}(energy)＝{エネルギー}.
CL_{EJ}(regulation)＝{規制}.
CL_{EJ}(protection)＝{保護}.

【０１０４】このとき、日本語概念集合{地球温暖化, オゾン層, 温室効果, フロン, 大気, 二酸化炭素, 環境}から英語概念集合への翻訳結果は{ozone layer, atmosphere･大気, carbon dioxide, environment, chrolofluorocarbon, warming}になる。

【０１０５】英語概念集合を構成する６つの英語概念のうち、「ozone layer」、「atmosphere･大気」、「carbon dioxide」及び「environment」の４つは、日本語概念集合中の日本語概念とシソーラス中で陽に結合されていたものである。また、「chrolofluorocarbon」及び「warming」の２つは、日本語概念集合中の日本語概念とシソーラス中で陽に結合されていなかったが、上記４つの英語概念の関連概念として追加されたものである。実は、「chrolofluorocarbon」は「フロン」の英語訳であり、「warming」は「地球温暖化」の英語訳の一部である。このようにして、概念結合として陽に表現されていない対訳を含む翻訳結果を得ることができる。

【０１０６】以上、この発明の実施の形態を図面を参照して詳述してきたが、具体的な構成はこれらの実施の形態に限られるものではなく、この発明の要旨を逸脱しない範囲の設計の変更等があってもよい。はじめに、上記実施の形態では、日本語と英語とのパラレルシソーラスを生成しているが、たとえば、日本語とフランス語、更には、日本語のものも含めて一般的な２ヶ国語によるパラレルシソーラスを生成するものであってもよい。

【０１０７】また、上記実施の形態では、図４に示したように、多義語と同義語の各々が有する概念を考慮した機能を実現しているが、多義語の概念のみを考慮した機能、又は同義語のみを考慮した機能を実現してもよい。この場合、日本語概念マージ３０７及び英語概念マージ３０８の機能を選択的に設けることで実現できる。

【０１０８】また、本発明におけるクライアント計算機２としては、有線回線により通信ネットワーク３に接続されるパーソナルコンピュータ又はワークステーション等、また、無線回線により通信ネットワーク３に接続される移動体通信端末（携帯電話、ＰＨＳ（Personal Handy-Phone System）、ＰＤＡ（Personal Digital Assistance）等）であってもよい。

【０１０９】なお、本発明のパラレルシソーラスの生成装置、及びパラレルシソーラスナビゲーションシステムは、このサーバ計算機１又はクライアント計算機２を機能させるためのプログラムによっても実現される。このプログラムは、たとえばＣＤ－ＲＯＭ等のコンピュータで読み取り可能な記録媒体に格納されている。

【０１１０】パラレルシソーラスの生成装置、又はパラレルシソーラスナビゲーションシステムを機能させるためのプログラムを記録した記録媒体は、図１に示す記憶装置１３そのものであってもよいし、また、外部記憶装置としてＣＤ－ＲＯＭドライブ等のプログラム読み取り装置（図示せず）が設けられ、そこに挿入することで読み取り可能なＣＤ－ＲＯＭ等であってもよい。また、上記記録媒体は、磁気テープ、カセットテープ、フロッピーディスク、ハードディスク、ＭＯ／ＭＤ／ＤＶＤ等、又は半導体メモリであってもよい。

【０１１１】また、本発明のパラレルシソーラスの生成装置により生成されたパラレルシソーラスは、ＣＤ－ＲＯＭ等のコンピュータで読み取り可能な記録媒体に格納されてもよい。このパラレルシソーラスは、二つの言語のシソーラスを結合したものであり、日本語及び英語概念ラベル生成３０２，３０３により二つの言語のタームを組み合わせた概念ラベルが生成され、日英概念結合３０４（図４参照）により、二つの言語の概念ラベルが概念に基づいて結合されている。

【０１１２】

【発明の効果】本発明のパラレルシソーラスの生成プログラムを記録した記録媒体によれば、二つの言語の関連シソーラスが結合されたパラレルシソーラスを二つの言語のテキストコーパスから自動的に生成することができる。生成されるパラレルシソーラスは、概念と概念の関連を示したものである。

【０１１３】また、本発明のパラレルシソーラスを記録した記録媒体によれば、タームとタームの関連を示す従来のシソーラスと異なり、多義語や同義語の問題が解決されているので、本発明により生成されたシソーラスを用いることで、各種の自然言語処理システムの精度を向上することができる。

【０１１４】また、本発明のパラレルシソーラスナビゲーションプログラムを記録した記録媒体によれば、複数言語にまたがる効率的なテキストマイニングが可能になる。特に、母国語のシソーラスをナビゲーションして外国語の情報にアクセスすることが容易になる。従来のクロスランゲージ情報検索において問題とされる検索要求の翻訳精度も、概念集合の遷移（翻訳）機能により大きく改善される。

【図面の簡単な説明】

【図１】本発明の実施の形態によるパラレルシソーラス

の生成装置と、該装置を収容するパラレルシソーラスナビゲーションシステムの構成を説明するブロック図である。

【図２】本発明の実施の形態によるパラレルシソーラスの生成装置における入出力とモジュール構成を機能的に説明する図である。

【図３】日本語シソーラス生成の処理の詳細を説明する図である。

【図４】日英シソーラス結合モジュールの処理の詳細を説明する図である。

【図５】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムにおいて、クライアント計算機の表示画面の内容を説明する図である。

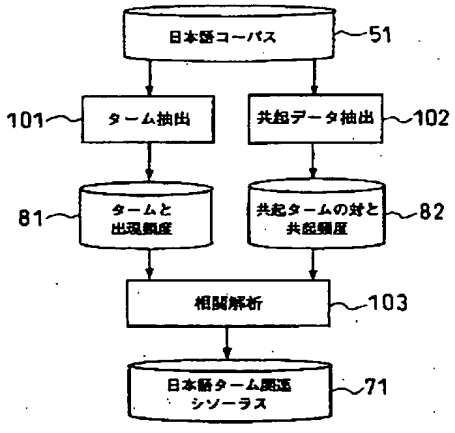
【図６】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムの処理を説明するフローチャートである。

【図７】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムを特徴付ける概念集合翻訳（ステップ４８０）の処理を詳細に説明する図である。

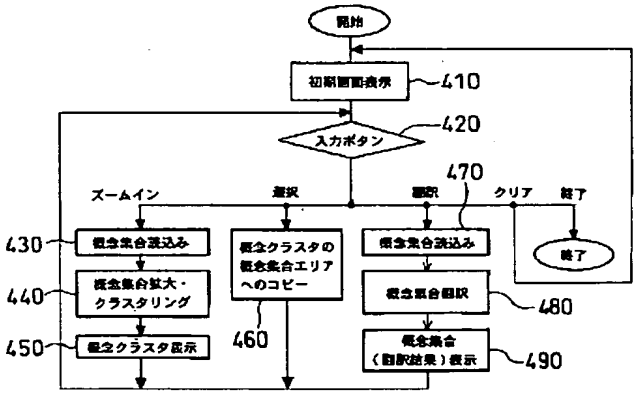
【符号の説明】

１　サーバ計算機
２　クライアント計算機
３　通信ネットワーク
１０　日本語シソーラス生成
１１　処理装置
１２　入力装置
１３　記憶装置
２０　英語シソーラス生成
３０　日英シソーラス結合
５１　日本語コーパス
５２　英語コーパス
６１　日本語概念関連シソーラス
６２　英語概念関連シソーラス
６３　日英概念結合データ
６４　英日概念結合データ
７１　日本語ターム関連シソーラス
７２　英語ターム関連シソーラス
７３　日英対訳辞書
７４　英日対訳辞書
８１　タームと出現頻度
８２　共起タームの対と共起頻度
９１　日英ターム結合データ
９２　英日ターム結合データ
９３　日本語概念ラベルデータ
９４　英語概念ラベルデータ
９５　日本語概念の関連タームデータ
９６　英語概念の関連タームデータ
１０１　ターム抽出
１０２　共起データ抽出
１０３　相関解析
３０１　日英ターム結合
３０２　日本語概念ラベル生成
３０３　英語概念ラベル生成
３０４　日英概念結合
３０５　日本語概念関連シソーラス生成
３０６　英語概念関連シソーラス生成
３０７　日本語概念マージ
３０８　英語概念マージ
１０１０　概念集合エリア
１０２０　ズームインエリア
１０２１　概念クラスタ
１０２２　選択ボタン
１０３０　ズームインボタン
１０４０　翻訳ボタン
１０５０　クリアボタン
１０６０　終了ボタン

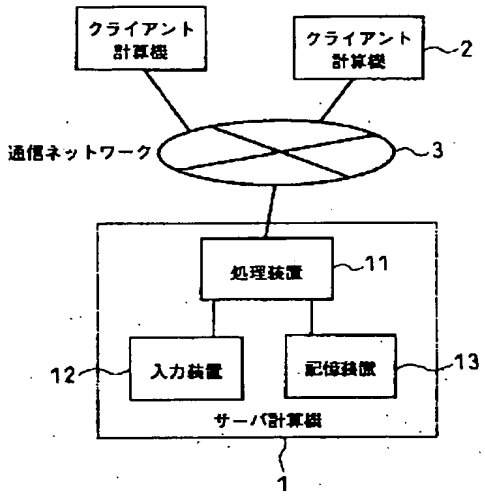
【図３】



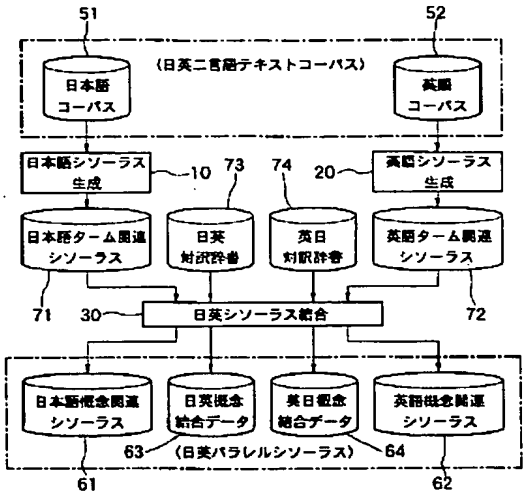
【図６】



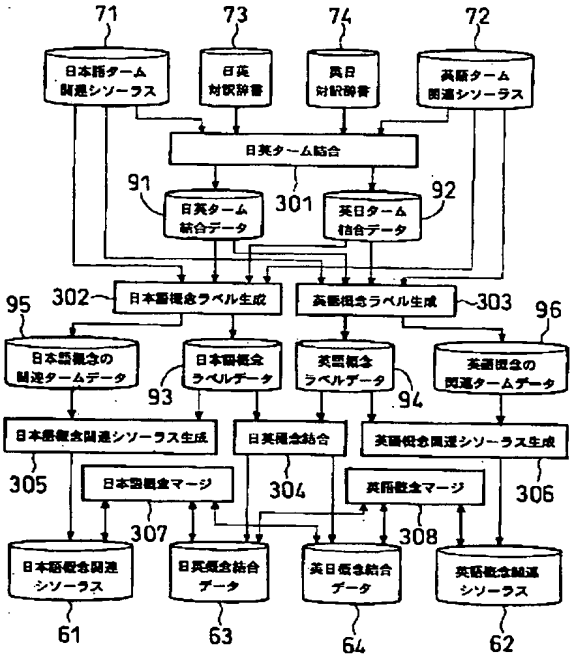
【図１】


クライアント計算機
クライアント計算機
2
通信ネットワーク
3
処理装置
11
入力装置
12
記憶装置
13
サーバ計算機
1

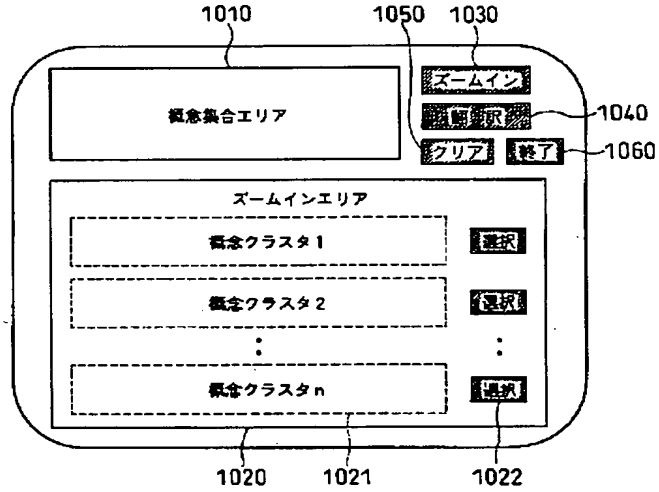
【図２】


51
52
(日英二言語テキストコーパス)
日本語コーパス
英語コーパス
日本語シソーラス生成
10
73
74
20
英語シソーラス生成
日本語ターム関連シソーラス
日英対訳辞書
英日対訳辞書
英語ターム関連シソーラス
71
30
日英シソーラス結合
72
日本語概念関連シソーラス
日英概念結合データ
英日概念結合データ
英語概念関連シソーラス
63
(日英パラレルシソーラス)
64
61
62

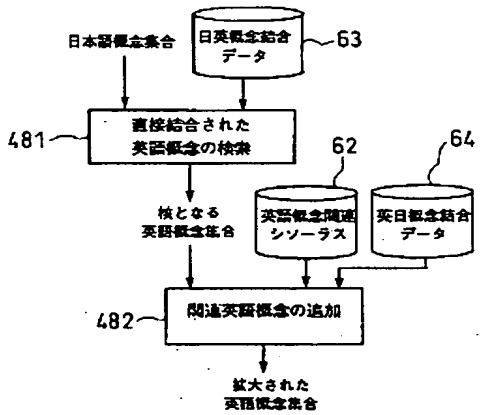
【図４】


71
73
74
72
日本語ターム関連シソーラス
日英対訳辞書
英日対訳辞書
英語ターム関連シソーラス
日英ターム結合
91
301
92
日英ターム結合データ
英日ターム結合データ
95
302
日本語概念ラベル生成
英語概念ラベル生成
303
96
日本語概念の関連タームデータ
日本語概念ラベルデータ
英語概念ラベルデータ
英語概念の関連タームデータ
93
94
日本語概念関連シソーラス生成
日英概念結合
英語概念関連シソーラス生成
305
日本語概念マージ
304
英語概念マージ
306
307
308
日本語概念関連シソーラス
日英概念結合データ
英日概念結合データ
英語概念関連シソーラス
61
63
64
62

【図５】


1010
1050
1030
概念集合エリア
ズームイン
選択解除
1040
クリア
終了
1060
ズームインエリア
概念クラスタ１
選択
概念クラスタ２
選択
概念クラスタｎ
選択
1020
1021
1022

【図７】


日本語概念集合
日英概念結合データ
63
481
直接結合された英語概念の検索
62
64
核となる英語概念集合
英語概念関連シソーラス
英日概念結合データ
482
関連英語概念の追加
拡大された英語概念集合